Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 6 | Issue : 3 | March 2020

RESEARCH ON CLOUD-BASED PLATFORM AND SERVICES FOR GENOMIC DATA ANALYSIS PIPELINES

Aditya Deshpande [1] | Swapnil Deshpande [2] | Rohan Kalgutkar [2] | Omkar Mahadar [2]

[1] Computer Department, Marathwada Mitra Mandal's COE, Pune, India. (IEEE Member)

[2] Computer Department, Marathwada Mitra Mandal's COE, Pune, India.

ABSTRACT

Bioinformatics administrations have been customarily given within the form of a web-server that is facilitated at organization foundation and serves numerous clients. The on demand, pay-as-you-go model creates a flexible and cost-effective means to access compute resources. This model, in any case, isn't adaptive enough to sufficiently manage with the expanding number of clients, expanding information measure, and modern necessities in terms of speed and accessibility of benefit. Genomic investigations incorporate an assortment of tools that address the worldwide changes of particular biological parameters. Genomic investigations that look at DNA, RNA, or protein levels give effective tools to characterize quality work. This paper provides information regarding facilitation of bio-analytic services using cloud-based platforms like AWS, Azure. These services involves tools like RNA-Seq Analysis Pipeline and Assembly Annotation Pipeline (Genome Assembly).

KEYWORDS: AWS, S3, EC2, EBS, RNA, Lambda, Pipeline, Genome.

I. INTRODUCTION:

Cloud:

Cloud computing is the on-demand delivery of IT resources over the Internet with pay-as-you-go pricing. Rather than purchasing, owning, and keeping up physical server farms and servers, you can get to innovation administrations, for example, processing force, stockpiling, and databases, dependent upon the situation from a cloud supplier like Amazon Web Services (AWS).

Amazon Web Services (AWS):

Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud platform, offering over 175 fully featured services from data centers globally.

Most functionality- AWS has significantly more services, and more features within those services, than any other cloud provider–from infrastructure technologies like compute, storage, and databases–to emerging technologies, such as machine learning and artificial intelligence, data lakes and analytics, and Internet of Things. This makes it faster, easier, and more cost effective to move the user's existing applications to the cloud and build nearly anything the user can imagine.
Most secure- AWS is architected to be the most flexible and secure cloud computing environment available today. Our core infrastructure is built to satisfy the security requirements for the military, global banks, and other high-sensitivity organizations. This is backed by a deep set of cloud security tools, with 230 security, compliance, and governance services and features.

Amazon Web Services includes functionalities like:

1. Amazon Elastic Compute Cloud
2. Amazon Elastic Block Store
3. Serverless Computing
4. Lambda functions
5. Batch Computing
6. Amazon Simple Storage Service
7. Web Server
8. Docker & Container

Comparison of various Cloud Computing platforms

Sr No.	Cloud Platform	Features	Cost	Description
1.	AWS	Highly Customizable	Free Trial	1. AWS was established in 2006.They gives on request distributed computing to huge IT Organizations 2. AWS is a cloud-based program for building business arrangements utilizing coordinated web administrations. 3. AWS incorporate Elastic Cloud Compute (EC2), Elastic Beanstalk, Simple Storage Service (S3) and Relational Database Service (RDS).
2.	Microsoft Azure	Windows and Linux Compatible	12 months free trial	Microsoft Azure was discharged about 10 years back, in 2010. Clients can run any help on the cloud or consolidate it with any current applications, server farm or foundation.
3.	Google Cloud	User Friendly	12 months free trial	1. Google Cloud Platform is Google's cloud specialist organization. 2. The stage empowers clients to make business arrangements utilizing Google-gave, particular web administrations.
4.	IBM Cloud	1. Pre-configured tools 2. Fully customizable management tools	-	IBM Cloud is a set of cloud computing services offered by the eponymous tech giant IBM. The solution offers platform as a service, software as a service and infrastructure as a service

II. TECHNIQUES AND IMPLEMENTATION METHODOLOGIES:

Amazon Web Services (AWS):

Cloud service allows enterprise class and individual users to acquire computing resources from large scale data centres of service providers. Users can rent machine instances with different capabilities as needed and pay at a certain per machine hour billing rate. Despite concerns about security and privacy, cloud service attracts much attention from both users and service providers.

A. Amazon Elastic Compute Cloud:

EC2 is a web service that provides resizable compute capacity in the cloud.

It is designed to make web-scale computing easier for developers. Amazon EC2 enables "compute" in the cloud.

Amazon EC2's simple web service interface allows the user to obtain and configure capacity with minimal friction.

It provides the user with complete control of the user's computing resources and lets the user run on Amazon's proven computing environment.

Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing the user to quickly scale capacity, both up and down, as the user's computing requirements change.

Amazon EC2 changes the economics of computing by allowing the user to pay only for capacity that the user actually use.

Instance Types:

1. Accelerated Computing instances
2. Compute Optimized instances
3. General Purpose instances
4. High Memory instances
5. Memory Optimized instances
6. Previous Generation instances
7. Storage Optimized instances

B. EBS:

Amazon Elastic Block Store (EBS) is an easy to use, high performance block storage service designed for use with Amazon Elastic Compute Cloud (EC2) for both throughput and transaction intensive workloads at any scale.

A broad range of workloads, such as relational and non-relational databases, enterprise applications, containerized applications, big data analytics engines, file systems, and media workflows are widely deployed on Amazon EBS. Data that is stored on an Amazon EBS volume will persist independently of the life of the instance.

Advantages of EBS:

EBS volumes are performant for the user's most demanding workloads, including mission-critical applications such as SAP, Oracle, and Microsoft products.

SSD-backed options include a volume designed for high performance applications and a general purpose volume that offers strong price/performance for most workloads.

HDD-backed volumes are designed for large, sequential workloads such as big data analytics engines, log processing, and data warehousing.

C. Serverless Computing:

Serverless computing allows the user to build and run applications and services without thinking about servers. With serverless computing, the user's application still runs on servers, but all the server management is done by AWS. At the core of serverless computing is AWS Lambda, which lets the user run his/her code without provisioning or managing servers.

D. Lambda:

AWS Lambda lets the user run code without provisioning or managing servers. The user pays only for the compute time the user consumes - there is no charge when the user's code is not running.

With Lambda, the user can run the code for virtually any type of application or backend service - all with zero administration.

The user can set up his/her code to automatically trigger from other AWS services or call it directly from any web or mobile app.

The appropriate time to use AWS Lambda and Ec2:

Amazon EC2 offers flexibility, with a wide range of instance types and the option to customize the operating system, network and security settings, and the entire software stack, allowing to easily move existing applications to the cloud. With Amazon EC2, the user is responsible for provisioning capacity, monitoring fleet health and performance, and designing for fault tolerance and scalability.

Whereas, AWS Lambda makes it easy to execute code in response to events, such as changes to Amazon S3 buckets, updates to an Amazon DynamoDB table, or custom events generated by the user's applications or devices.

AWS Lambda functions:

The code any user runs on AWS Lambda is uploaded as a "Lambda function". Each function has associated configuration information, such as its name, description, entry point, and resource requirements.

The code must be written in a "stateless" style i.e. it should assume there is no affinity to the underlying compute infrastructure.

E. AWS Batch:

AWS Batch is a set of batch management capabilities that enables developers, scientists, and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS.

AWS Batch dynamically provisions the optimal quantity and type of compute resources (e.g., CPU or memory optimized instances) based on the volume and specific resource requirements of the batch jobs submitted.

With AWS Batch, there is no need to install and manage batch computing software or server clusters, allowing the user to instead focus on analysing results and solving problems.

F. Batch Computing:

Batch computing is the execution of a series of programs ("jobs") on one or more computers without manual intervention. Input parameters are predefined through scripts, command-line arguments, control files, or job control language.

A given batch job may depend on the completion of preceding jobs, or on the availability of certain inputs, making the sequencing and scheduling of multiple jobs important, and incompatible with interactive processing.

Benefits of batch computing:

1. It can shift the time of job processing to periods when greater or less expensive capacity is available.
2. It avoids idling compute resources with frequent manual intervention and supervision.
3. It increases efficiency by driving higher utilization of compute resources.
4. It enables the prioritization of jobs, aligning resource allocation with business objectives.

G. Amazon Simple Storage Service (S3):

Amazon S3 is object storage built to store and retrieve any amount of data from anywhere on the Internet.

It's a simple storage service that offers an extremely durable, highly available, and infinitely scalable data storage infrastructure at very low costs.

Benefits of S3:

Amazon S3 provides a simple web service interface that the user can use to store and retrieve any amount of data, at any time, from anywhere on the web. Using this web service, the user can easily build applications that make use of Internet storage.

Since Amazon S3 is highly scalable and the user only pays for what he/she uses, the user can start small and grow his/her application as per the requirement, with no compromise on performance or reliability.

Features provided by S3:

Amazon S3 enables any developer to leverage Amazon's own benefits of massive scale with no up-front investment or performance compromises.

Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure their data is quickly accessible, always available, and secure.

H. Web Server:

A web server is server software, or hardware dedicated to running said software, that can satisfy World Wide Web client requests. A web server can, in

general, contain one or more websites. A web server processes incoming network requests over HTTP and several other related protocols.

The primary function of a web server is to store, process and deliver web pages to clients. The communication between client and server takes place using the Hypertext Transfer Protocol (HTTP). Pages delivered are most frequently HTML documents, which may include images, style sheets and scripts in addition to the text content.

Market share as of July 2018		
Product	**Vendor**	**Percent**
Apache	Apache	44.3%
Nginx	NGINX, Inc.	41.0%
IIS	Microsoft	8.9%
LiteSpeed WebServer	LiteSpeed Technologies	3.9%
GWS	Google	0.9%

I. Docker:

Docker is a set of platform as a service (PaaS) products that use OS-level virtualization to deliver software in packages called containers.

Containers are isolated from one another and bundle their own software, libraries and configuration files; they can communicate with each other through well-defined channels. All containers are run by a single operating-system kernel and are thus more lightweight than virtual machines.

Docker image:

A Docker image is a file, comprised of multiple layers, used to execute code in a Docker container.

An image is essentially built from the instructions for a complete and executable version of an application, which relies on the host OS kernel.

Some essential docker commands:

1. Build- Docker build is a command used to build an image from a dockerfile.

2. Syntax- docker build [OPTIONS] IMAGE .

   . (dot) indicates to take the dockerfile from the current directory.

3. Run- Docker run is used to run a docker image and create an image container.

   Syntax- docker run [OPTIONS] IMAGE [COMMAND] [ARG...]

4. Execute- Docker execute is used to run a command in an existing container.

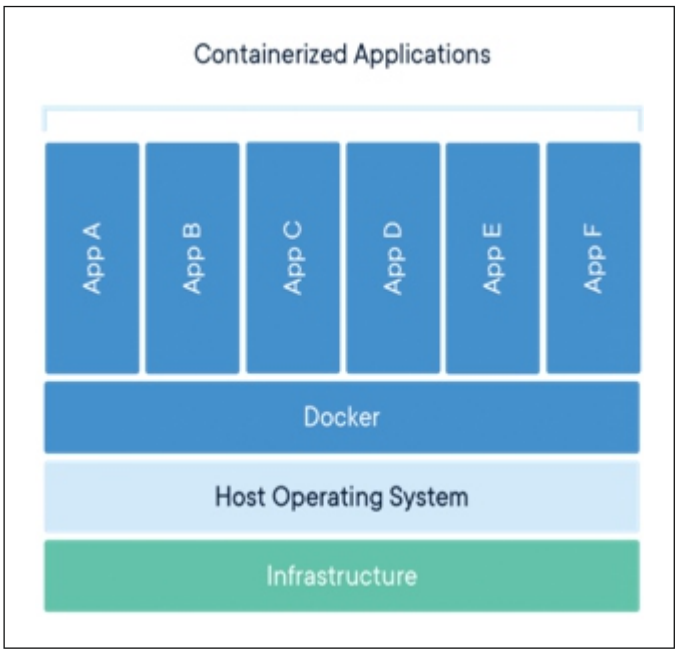
   Syntax- docker exec [OPTIONS] CONTAINER COMMAND [ARG...]

Container:

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.

Container images become containers at runtime and in the case of Docker containers - images become containers when they run on Docker Engine. Available for both Linux and Windows-based applications, containerized software will always run the same, regardless of the infrastructure.

Containers isolate software from its environment and ensure that it works uniformly despite differences for instance between development and staging.



III. CONCLUSION:

The objective of this research paper is to study various methodologies and concepts to be implemented for successful completion of our product. From our investigation it is concluded that AWS batch along with docker will be used for processing of the genomics data. We are also going to use AWS EC2 instance on which the host website will be services to the user where the required file will get uploaded for genomic processing. The file when gets uploaded to the S3 storage will trigger the lambda function which will activate AWS batch processing. AWS Batch will implement the genomic analysis pipelines like RNA-Seq Analysis Pipeline and Assembly-Annotation pipeline. The result from these pipelines will be stored back into S3 and will be made available for user to download.

IV. ACKNOWLEDGEMENT:

We acknowledge project guide, Dr. H.K. Khanuja, Head Of Department, Computer Engineering Department, Marathwada Mitra Mandal's College Of Engineering Pune and Mr. Mandar Sahasrabuddhe, Business Head, Kovid BioAnalytics, for providing us with internship and experience in live project and team coordination.

REFERENCES:

1. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome
2. Pearson, S., Benameur, A., Privacy, Security and Trust Issues Arises from Cloud Computing, Cloud Computing Technology and Science (CloudCom), IEEE Second International Conference,2010, On page(s): 693-702.
3. Puneet Jai Kaur, Sakshi Kaushal, Security Concerns in Cloud Computing, Communication in Computer and Information Science Volume 169 in 2011, On page(s): 103-112.
4. M. Palankar, A. Iamnitchi, M. Ripeanu, and S. Garfinkel, "Amazon S3 for science grids: a viable solution?" in Proceedings of the 2008 international workshop on Data-aware distributed computing. ACM, 2008, pp. 55–64.
5. C. Evangelinos and C. Hill, "Cloud Computing for parallel Scientific HPC Applications: Feasibility of running Coupled Atmosphere-Ocean Climate Models on Amazons Ec2." ratio, vol. 2, no. 2.40, pp. 2–34, 2008
6. Hongkun Yang, Simon S. Lam," Real-Time Verification of Network Properties Using Atomic Predicates", IEEE/ACM Trans. Netw. 2016
7. Karthick Jayaraman, Nikolaj Bjørner, Geoff Outhred, Charlie Kaufman," Automated Analysis and Debugging of Network Connectivity Policies", 2014
8. Azure.microsoft.com. (2020). What is cloud computing? A beginner's guide | Microsoft Azure. [online] Available at: https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/ [Accessed 15 Jan. 2020].
9. Amazon Web Services, Inc. (2020). What is Cloud Computing. [online] Available at: https://aws.amazon.com/what-is-cloud-computing [Accessed 15 Jan. 2020].